Project 1

Clara Torslov (ct32699)

Part 1

Question: Which sports have the tallest or shortest athletes? And does the distribution of heights change for the various sports between medalists and non-medalists?

Introduction: We are working with the dataset, `olympics_top`, containing historical data of the Olympic Games, from Athens 1896 to Rio 2016. This dataset contains the same columns as the `olympics` dataset, plus the additional columns: `decade`, `gold`, `medalist` and `medal`. Each row in the dataset corresponds to a competiter in the olympics and the columns represent different information about that competitor.

Throughout part 1 we consider the variables `height`, `sport` and `medalist`. Height states the height of the given athlete as measured in inches. Sport states which sport the athlete competes in independent of events, fx. `Speed Skating Women's 500 metres` and `Speed Skating Women's 1,000 metre` are both `Speed Skating`. Medalist states whether an athlete won a medal or no medal, independent of which medal the athlete won. The variable takes the values "medalist" or "no medal".

Approach: To answer the first part of this question we will plot the distribution of sport vs the athletes height. We would like to properly visualize the distribution of height among the different sports and further be able to compare. We use boxplots (`geom_boxplot()`) as they are useful when visualizing and comparing multiple distributions. To better compare we order the boxplots by the median (`reorder(sport, height, median, na.rm = TRUE)`). We see that our data has missing height values (`table(is.na(olympics_top$height))`) and have therefore added `na.rm = TRUE` to remove missing values before ordering.

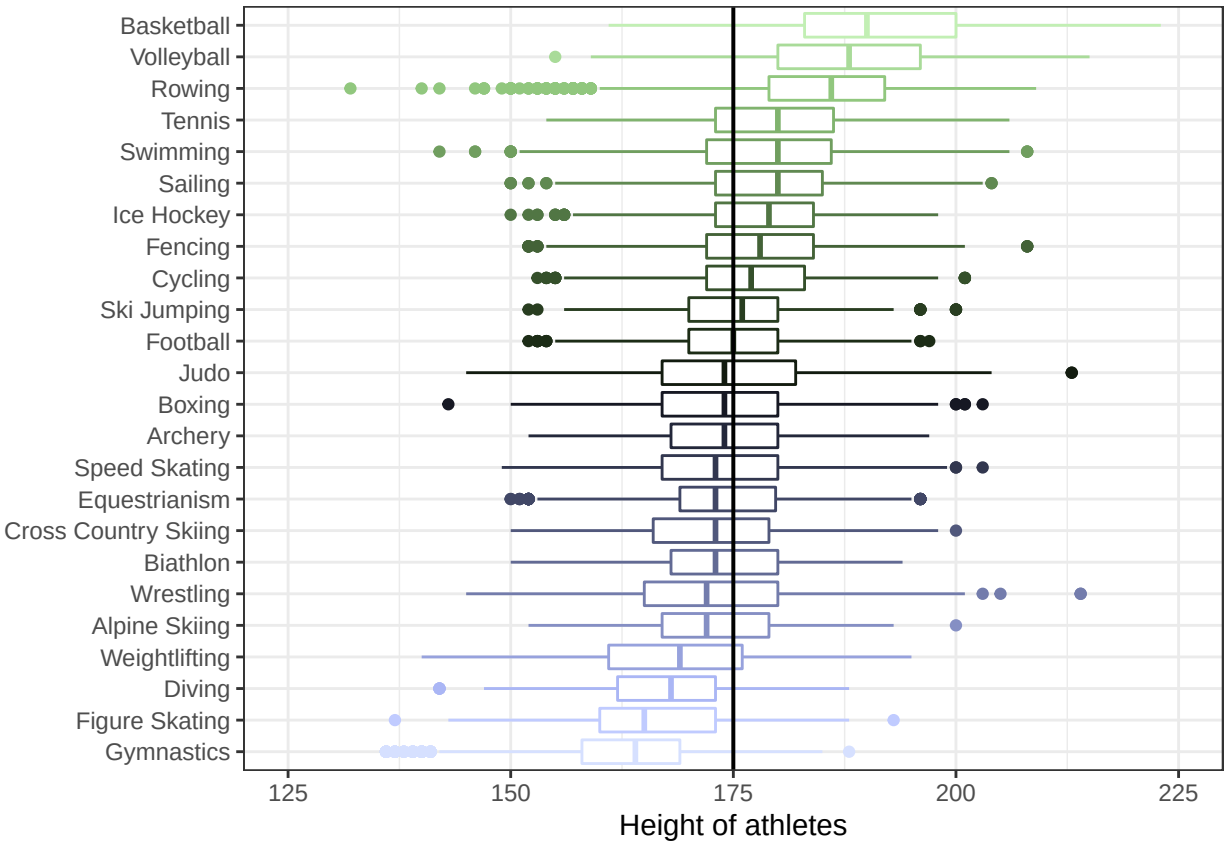
We further wish to examine if the distribution of heights change for the various sports between medalists and non-medalists. To examine this we use ridgeline plots, as they are often useful when visualizing and comparing a larger number of distributions across several groups. Further ridgeline plots will accurately represent bimodal data whereas a boxplot will not.

Analysis: First we plot the distribution of sport vs height. We color by sport. A black line indicating the median height across all sports have been added.

```
#table(is.na(olympics_top$height))

olympics_top %>%
ggplot() +
aes(y = reorder(sport, height, median, na.rm = TRUE),
    x = height,
    color = reorder(sport, height, median, na.rm = TRUE)
    ) +
geom_boxplot(show.legend = FALSE,
             na.rm = TRUE) +
theme_bw() +
scale_y_discrete(name = NULL) +
scale_x_continuous(name = "Height of athletes",
                   breaks = c(125,150,175,200,225),
                   labels = c("125","150","175","200","225"),
                   limits = c(125,225)
```

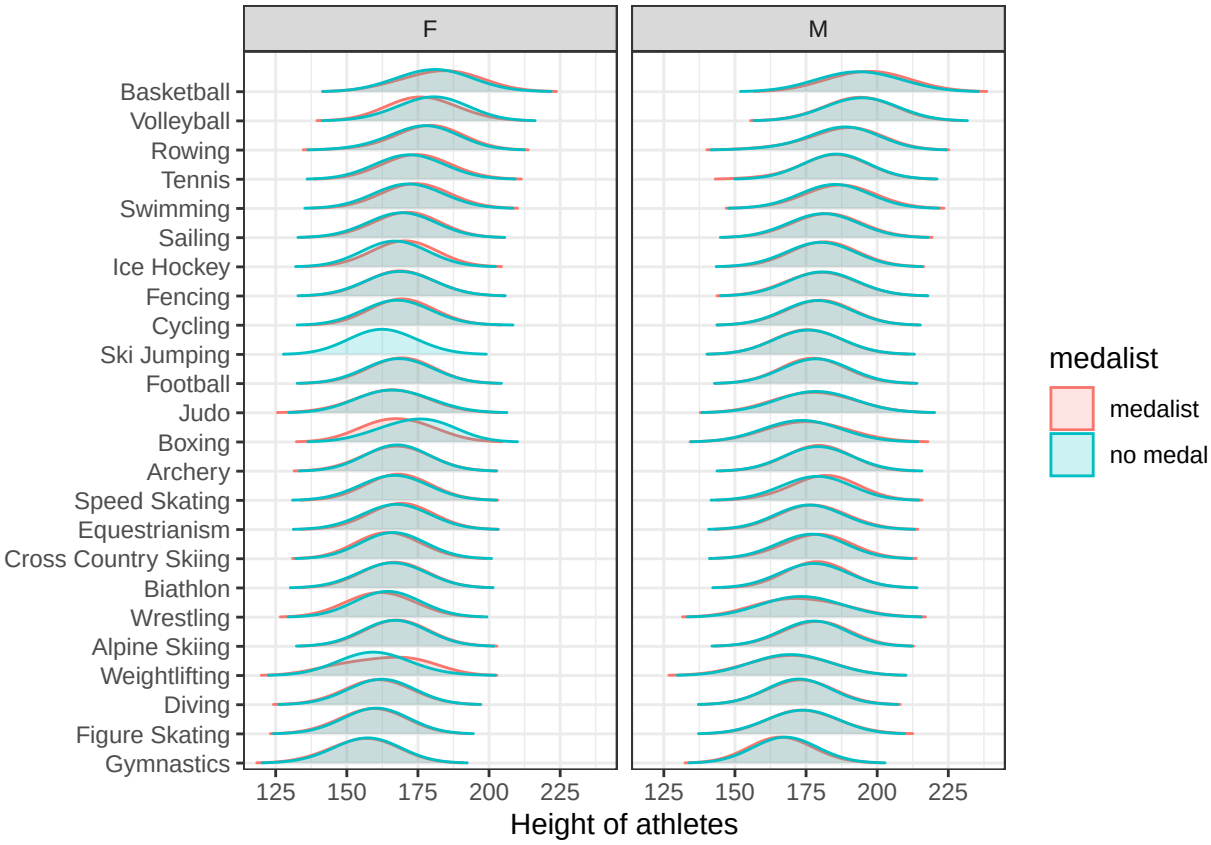
```
                    ) +
  geom_vline(xintercept = median(olympics_top$height, na.rm = TRUE),
             color = "black",
             size = 0.7) +
  scale_color_discrete_diverging(palette = "Tofino")
```



We plot the distribution of sport vs height between medalists and non-medalists, using ridgeline plots. We use `rel_min_height = 0.01` to better visualize the tails. As women and men have different height it makes sense to facet across sex.

```
ggplot(olympics_top) +
aes(y = reorder(sport, height, median, na.rm = TRUE),
    x = height,
    fill = medalist,
    color = medalist) +
geom_density_ridges(alpha = 0.2,
                    rel_min_height = 0.01,
                    scale = 0.9,
                    bandwidth = 10,
                    na.rm = TRUE) +
theme_bw() +
scale_y_discrete(name = NULL,
                 expand = expansion(mult = c(0.01, 0.06))) +
scale_x_continuous(name = "Height of athletes",
                   breaks = seq(125, 225, by = 25),
                   labels = c("125","150","175","200","225"),
                   limits = c(115,245),
                   expand = expansion(mult = c(0.01, 0.0))
```

```
                    ) +
  facet_wrap(vars(sex))
```



We see one of the distributions are missing and examine the group.

```
olympics_top %>%
  filter(sport == "Ski Jumping" & sex == "F" & medalist == "medalist")
```

```
## # A tibble: 2 x 18
##       id name  sex     age height weight team  noc   games  year season city
##    <dbl> <chr> <chr> <dbl>  <dbl>  <dbl> <chr> <chr> <chr> <dbl> <chr>  <chr>
## 1  76645 Coli~ F        18    165     57 Fran~ FRA   2014~  2014 Winter Sochi
## 2 127078 Cari~ F        22    171     62 Germ~ GER   2014~  2014 Winter Sochi
## # ... with 6 more variables: sport <chr>, event <chr>, medal <chr>, gold <chr>,
## #   medalist <chr>, decade <dbl>
```

Discussion: Basketball and Volleyball have the tallest athletes. This makes sense when considering these two sports are very dependent upon the height of the athletes. The fact that these sports have the tallest athletes is made very clear by the ordered boxplots. There are 10 sports in total, which has a median height above the average median height across all athletes, and 13 sports below the general median. This indicates that the sports with a median above the general median has very many very tall athletes, thereby pulling the general median up. The sport with the shortest athletes is Gymnastics, which is made clear as the third quantile of the boxplot is well below the overall median. This is also the case for Figure Skating and Diving.

When considering the ridgeline plots, at first glance it does not look like the distribution changes significantly for the various sports between medalists and non-medalists. This is seen as the distributions have the same shape and overlap alot (grey areas). However looking more closely we see some smaller difference (i.e. color). Fx. in Basketball the distribution of height is clearly shifted slightly to the right, for both male

and female, indicating taller medalist athletes. Generally, for the female athletes, when moving from tallest to shortest athletes, the red tail is visible to the right (indicating taller medalist athletes) and shifting to being visible to the left (indicating shorter medalist athletes) for the shortest sports. For the males it looks as if there is no significant difference of height except for minor (probably insignificant) differences in the tails. However we would have to run statistical analysis to determine if any of above observations are statistically significant. We see a noticeable difference in two sports, namely womens boxing and womens weightlifting, where non-medalists and medalists seem to be shorter and taller respectively. This is indicated by a shifted distribution.

Note: the 'females medalist Ski Jumping'-group only has two observations as it's a new sport (first season was 2014 Winter) and therefore not enough to constitute a distribution.

Part 2

Question: Has the number of athletes and ratio of females vs males changed over time? Is it the same pattern across all participating countries?.

Introduction: We are still working with the `olympics_top` dataset. In this problem we consider the variables `sex`, `decade` and `team`. `sex` states the sex of a given athlete as measured by "F" (female) or "M" (male). `team` states what country a given athlete competes for among the 14 different countries included in the dataset. `decade` is an added column made from the `year` variable and states which decade the given athlete competed in the Olympics.

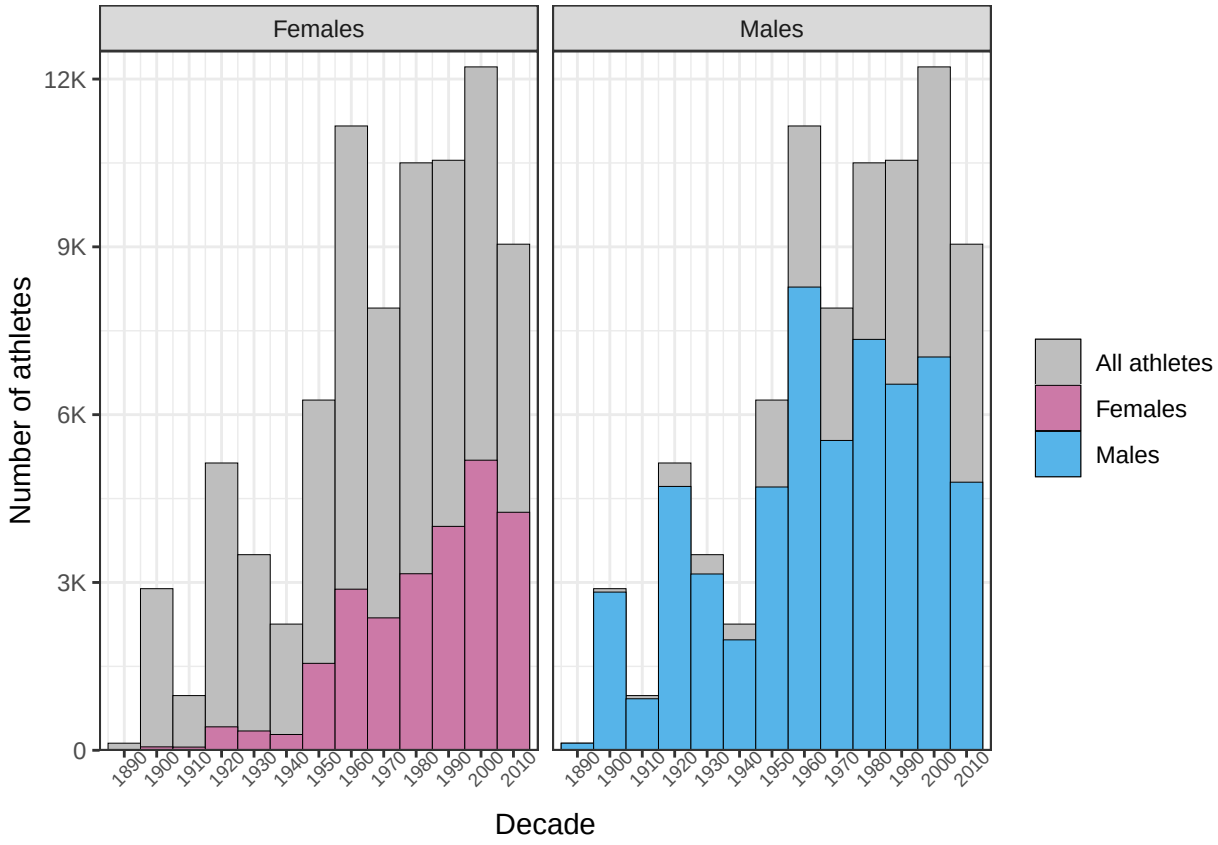
Note that each row in the dataset corresponds to a competiter in the olympics. A specific athlete can be represented several times, as the same athlete can compete in multiple events (fx. in both 500m and 1000m in the same sport). This is not a concern to us as we wish to count number of spots/competitors rather than specific athletes, meaning if one athlete did not take a double spot another athlete would occupy one of the spots in the competition. So, when we later write 'all athletes' we therefore refer to number of competitor spots.

Approach: To answer the question posed in the second part, we first wish to plot the distribution of all athletes over time. We wish to do so while simulaneously plotting the distribution of female and male athletes seperately. We wish to do so by faceting across `sex` and using a bar plot (`geom_bar()`), as they are great for visualizing larger changes over time and further as the data is not aggregated beforehand. Further we wish to examine the female vs. male proportion with a stacked density plot, as it is great for visual representation of changing proportions over time. Further we have a continuous time variable (`year`) and a large dataset. We need both plots to answer the first question in part 2. To further examine if the pattern we may or may not find in the male/female ratio we facet the stacked density plot across countries.

Analysis: We first do a bar plot of all athletes and one sex over time with sex faceted.

```
olympics_top %>%
  mutate(sex = recode(sex, "F" = "Females", "M" = "Males")) %>%
ggplot() +
  aes(x = decade) +
      geom_bar(
    data = select(olympics_top, decade),
    color = "black",
    aes(fill = "All athletes"),
    size = 0.1,
    width = 10
  ) +
  facet_wrap(vars(sex)) +
  geom_bar(aes(fill = sex),
           color = "black",
           size = 0.1,
           width = 10) +
```

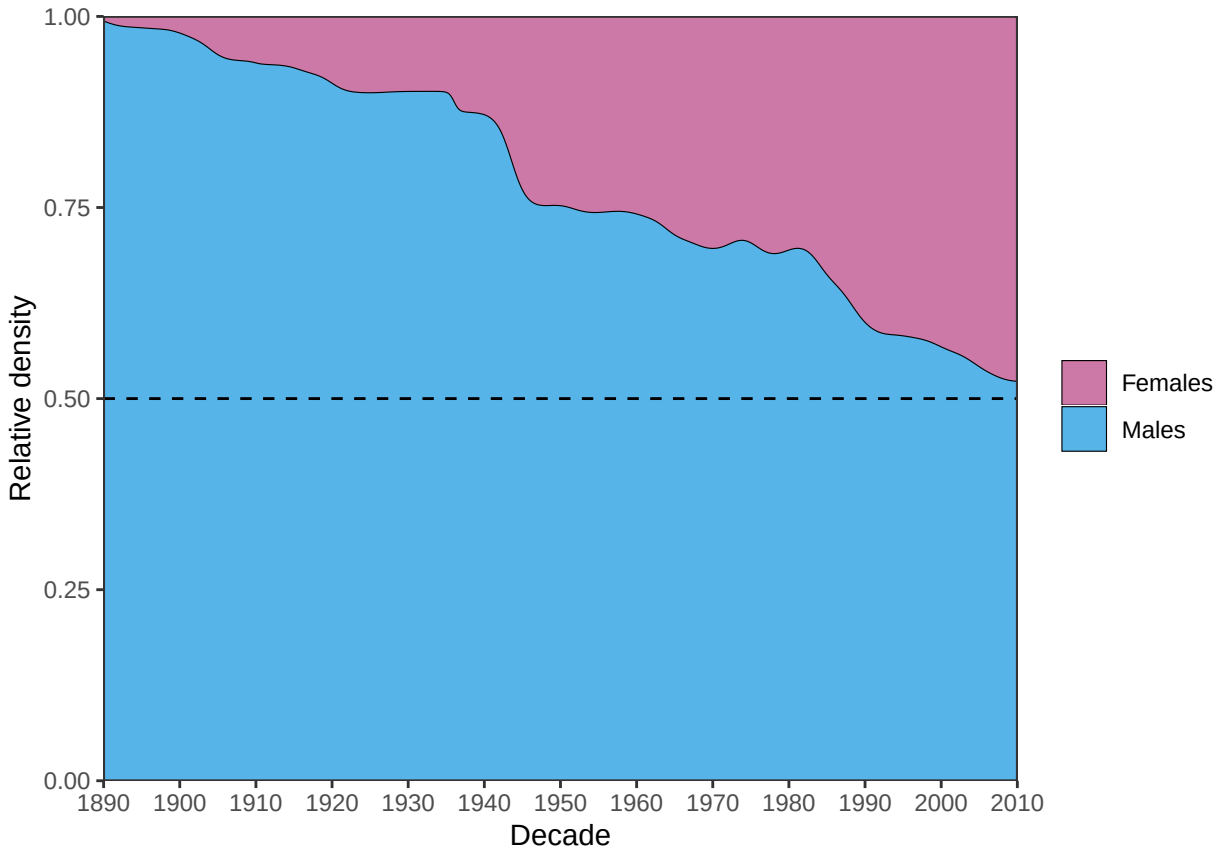
```
  theme_bw() +
  scale_x_continuous(name = "Decade",
                     breaks = seq(1890, 2010, by = 10),
                     labels = c("1890","1900","1910","1920","1930","1940",
                                "1950","1960","1970","1980","1990","2000","2010"),
                     expand = c(0.02, 0.01)) +
  scale_y_continuous(name = "Number of athletes",
                     expand = c(0, 0),
                     limits = c(0,12500),
                     breaks = seq(0,12000, by = 3000),
                     labels = c("0", "3K", "6K", "9K", "12K")
                     ) +
  labs(fill = NULL) +
  theme(axis.text.x=element_text(size=rel(0.8), angle=45) ) +
  scale_fill_manual(values = c("All athletes" = "grey",
                               "Females" = "#CC79A7",
                               "Males" = "#56B4E9"))
```



We plot the distribution of the two sexes against eachother. We choose a smaller bandwidth for the density to more clearly see the difference between males and females rather than the general overall distribution. Note that we now plot `year`, as it's a continuous variable, however still label the x-axis by decade.

```
olympics_top %>%
  mutate(sex = recode(sex, "F" = "Females", "M" = "Males")) %>%
  ggplot() +
  aes(x = year,
      y = after_stat(count),
```

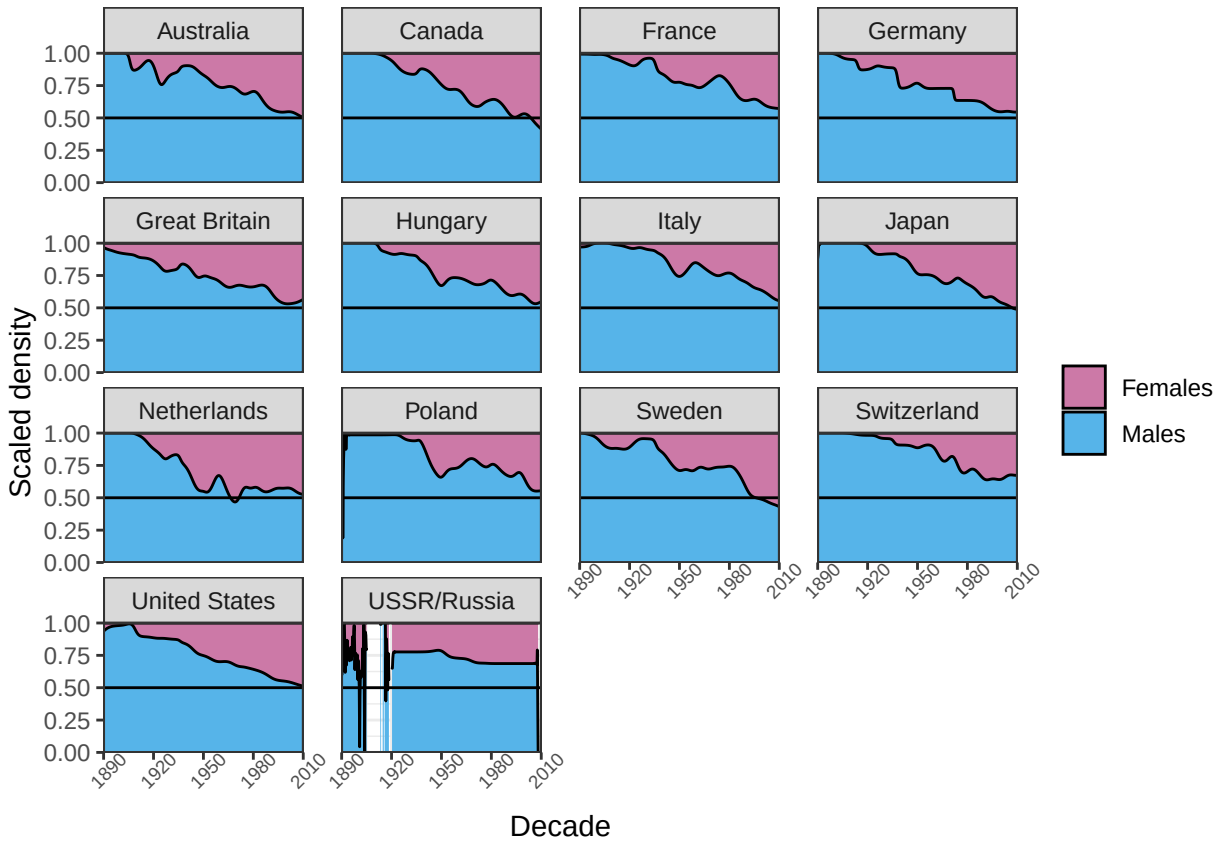
```
      fill = sex) +
  geom_density(bw = 2,
               position = "fill",
               size = 0.2) +
  geom_hline(yintercept = 0.5, linetype = "dashed") +
  scale_x_continuous(name = "Decade",
    breaks = seq(1896, 2016, by = 10),
    labels = c("1890","1900","1910","1920","1930","1940",
               "1950","1960","1970","1980","1990","2000","2010"),
    expand = c(0, 0)) +
  scale_y_continuous(name = "Relative density",
                     expand = c(0, 0)) +
  theme_bw() +
  labs(fill = NULL) +
  scale_fill_manual(values = c("Females" = "#CC79A7", "Males" = "#56B4E9"))
```



Lastly, we plot the distribution of the two sexes against eachother faceted across country.

```
olympics_top %>%
  mutate(sex = recode(sex, "F" = "Females", "M" = "Males")) %>%
ggplot() +
  aes(x = year,
      y = after_stat(count),
      fill = sex) +
geom_density(bw = 3,
             position = "fill") +
geom_hline(yintercept = 0.5) +
  scale_x_continuous(name = "Decade",
```

```
                    breaks = seq(1896, 2016, by = 30),
                    labels = c("1890", "1920","1950", "1980", "2010"),
                    expand = c(0,0)
                  ) +
scale_y_continuous(name = "Scaled density",
                    expand = c(0, 0)) +
theme_bw() +
theme(axis.text.x=element_text(size=rel(0.8), angle=45),
      panel.spacing.x = unit(1, "lines")) +
labs(fill = NULL) +
scale_fill_manual(values = c("Females" = "#CC79A7", "Males" = "#56B4E9")) +
facet_wrap(vars(team))
```



Discussion: In general we see an increase of athletes participating as time increases, with few dips due to historical or political reasons. In the bar plot these decreases in athlete participation (grey bars) can be clearly seen from 1900 to 1910 (World War I), from 1920 to 1930 and down again in 1940 (due to World War II and the Great Depression) and 1960 to 1970 (apartheid boycott). In the Olympics' first 6 decades it was primarily men who participated. This is seen as the blue bars have approximately the same height as the grey bars. Moving further along in time it can be seen from the rising red bars that female participation is increasing, however so is the total number of participants. From the second plot we see how the proportion of females have increased throughout time to a point where the overall ratio of males and females is almost equal. The largest increase in the female/male ratio is seen around 1945 - during/after World War II where many men were affected by the war.

The pattern of increase in the porportion of females is seen throughout all the participating countries (of the data set). Although overall similar, the pattern still differs among the individual countries. Canada and Sweden stands out as the most inclusive countries, with the female sex in most recent decades being

predominant. Least inclusive is Switzerland which in most recent decades sees a decline in female participants set against male participants. This is also seen in Great Britain, as the pink density decreases in the last decade. Note that the graph of USSR/Russia looks weird. A quick bar plot will show the reader that they only participated in a few decades and thus the distribution is smoothed in the missing years.